Survey Research Methods (2011)
Vol.5, No.3, pp. 89-102
ISSN 1864-3361
http://www.surveymethods.org

The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency

Carsten Sauer
Bielefeld University

Katrin Auspurg
University of Konstanz

Thomas Hinz
University of Konstanz

Stefan Liebig
Bielefeld University

Over the last decade, there has been a marked increase in the number of studies on attitude and decision research which use the factorial survey (FS) design. The FS integrates experimental set-ups into a survey: respondents react to hypothetical descriptions (vignettes) while the values of each attribute (dimension) of these descriptions systematically vary in order to estimate their impact on respondent judgments. As the vignettes are based on a number of dimensions and as respondents evaluate several vignettes, FSs are demanding in terms of individual cognitive and information-processing abilities. So far, there is little empirical knowledge of whether and to what extent this complexity is feasible in general population samples with heterogeneous respondents. Using data from a study on the fairness of earnings (with a mixed mode sample consisting a computer assisted personal interview [CAPI], computer assisted self interview [CASI], and paper and pencil [PAPI] mode), the complexity of FSs is analyzed in terms of: 1) design dimensions, such as the number of vignette dimensions (five, eight, or 12) and the number of vignettes for single respondents (10, 20, or 30), which were varied in a 3x3 experimental design; and 2) respondent characteristics that are associated with cognitive abilities (age and education). Two different indicators for cognitive load as well as learning and fatigue effects are analyzed: 1) latency time and 2) response consistency. The results show that raw reaction times but not latency times are longer for older respondents, suggesting that the cognitive effort needed for the evaluation of vignettes is not particularly high. Consistency measures reveal that respondents with a lower educational level show greater inconsistency in their evaluations when the number of vignettes is high. The number of dimensions has an effect on consistency only when respondents have to rate a large number of vignettes. In short, the results demonstrate that FSs are applicable in general population samples but should be used with a limited number of vignettes and dimensions per respondent.
Keywords: factorial survey, vignettes, response time, response consistency, cognitive load

1 Introduction

Over the last decade there has been a marked increase in the number of studies on attitude and decision research using the FS design (Wallander 2009).[1] This method was first established in the social sciences by Peter H. Rossi in his dissertation of 1951. It was used for the measurement of social status and household prestige (Alves and Rossi 1978; Rossi 1979). Rossi's idea was to develop a technique to measure the (relative) impact of causal factors on social attitudes and therewith to determine judgment principles (Rossi and Anderson 1982).

A FS is an experimental method that confronts respondents with several hypothetical descriptions of objects or situations (*vignettes*). Within these descriptions, the attributes (*dimensions*) vary systematically in their values (*levels*) (Jasso 2006). The respondents' task is to express their evaluations or to decide how they would act in the situations presented. The aim is to identify how each dimension affects these evaluations or decisions and to assess the relative impact of the single dimensions.

Despite the growing number of applications of the FS design in social research, there is little empirical knowledge on its methodological implications. This is especially true for the use of FSs in general population samples. Many previous studies have used homogeneous respondent populations, most often students, and were carried out in the lab or in comparable settings (e.g. Jasso and Meyersson Milgrom 2008; Jasso and Webster 1999; for an overview: Wallander 2009).

Contact information: Carsten Sauer, Bielefeld University, P.O. 100131, D-33501 Bielefeld, e-mail: carsten.sauer@uni-bielefeld.de

[1] This paper is based on a research project funded by the DFG (German Research Foundation) at the universities of Konstanz and Bielefeld. The project is part of the DFG Priority Program "Survey Methodology". For more details: http://www.soziologie.uni-konstanz.de/professuren/prof-dr-thomas-hinz/forschung/aktuelle-forschungsprojekte/fs0/factorial-survey/

As the research design can be complex and the respondents' rating task in FSs demanding, a number of method effects may occur and cause measurement errors and misleading interpretations of results. Those method effects are to be expected especially in general population samples where respondents are more heterogeneous in terms of cognitive abilities and the ability to focus on a single task compared to student samples (cf. Henrich, Heine and Norenzayan 2010).

This paper addresses the question of how respondents differing in age and educational background cope with the complex evaluation task in FSs. Data stem from a survey dealing with the perceived fairness of earnings. A complexity variation for the method was created by randomly allocating respondents to questionnaire splits differing in the number of vignette dimensions (five, eight or 12) and the number of vignettes (10, 20 or 30) for single respondents. Response times and response consistency – both known indicators for response problems – are analyzed (see for response times: Bassili and Scott 1996; Malhotra 2008; for response consistency: Finucane, Slovic and Schmidt 2005; Swait and Adamowicz 2001).[2] Additionally, the possible learning and fatigue effects are studied. The main research questions are as follows: 1) At which level of complexity do FSs provoke a high or too high cognitive load for respondents (as signalled by long processing times, fatigue effects and inconsistent responses)? 2) Do older or less educated respondents have more problems when evaluating (complex) FSs, leading to the indicated symptoms of a high cognitive (over)load?

2 Cognitive Load in Factorial Surveys

The evaluation task in FSs can be complex when respondents have to cope with a high amount of information. The complexity stems mainly from two different sources: the *number of dimensions* used to describe the vignettes and the *number of vignettes* presented to single respondents. The higher the number of dimensions, and the more vignettes presented, the more demanding the evaluation task. Respondents with higher cognitive abilities – e.g. students – are expected to perform relatively well when evaluating vignettes. As long as FSs are used within homogeneous student samples, the high degree of complexity should not cause a major problem since students are in most cases used to dealing with complex tasks. In psychology and economics, students regularly take part in laboratory experiments and surveys and seem to be able to cope with a high demand for cognitive attention. Measurements from FSs might, however, be more problematic if the respondents are not generally accustomed to tasks with high cognitive demands and if respondents show a high variability in cognitive abilities. Therefore, the application of FSs in general population samples requires more methodological attention.

In population samples, a high heterogeneity of respondents in terms of age and educational background exists, and both characteristics are correlated with cognitive performance. Previous research has shown that older people tend to have more difficulties than younger people in retaining new information in the working memory; this is especially true in regard to written information (Krosnick 1991; Radvansky and Copeland 2004). Copeland and Radvansky (2007) found that, compared to older people, younger people read more quickly and assimilated information more easily.[3]

Apart from age, cognitive ability is obviously related to educational achievement and formal levels of schooling (Falch and Sandgren 2006) because the hierarchical stratification of educational levels goes hand-in-hand with the varying and prolonged stimulation of cognitive ability, and also because the selection processes within the education system are at least partly based on cognitive performance (Brody 1997). Previous research has already shown that the educational background of respondents influences the occurrence of method effects like context or order effects in attitude surveys (Krosnick 1991; Krosnick and Alwin 1987; Narayan and Krosnick 1996; Sniderman and Grob 1996).

The results on response effects by both age and educational background have direct implications for the feasibility of FSs in general population samples. Most empirical applications of FSs present vignettes in the form of written text.[4] Complexity increases by the amount of text per vignette (the number of dimensions) and the number of vignettes presented. In most cases, the respondents are asked to give a rating or evaluation for each vignette after reading. It can lead to substantially wrong conclusions if especially older or less well-educated people have difficulties in processing the information and fulfilling the rating task. For instance, if data analysis revealed differences in the influence of dimensions by respondent age and educational background, this might not reflect substantial differences in their attitudes but rather in their differing capabilities for task-handling. It is therefore necessary to know whether there is any evidence for the existence of information overload, especially of older and less educated respondents.[5]

Based on the findings of previous research on response behavior in general population samples, this study focuses on the effects of complexity of vignettes and of age and educational background of the respondents on two methodological signals for a high cognitive load and possible problems

[2] Response times that are needed to finish FS modules of different complexity are additionally informative for assessing the practical feasibility of FSs.

[3] In their experiment, respondents had to combine the information of three sentences about the spatial relation of two objects into a single spatial arrangement. Significant differences in the performance of this task between younger and older respondents were detected: the ability to comprehend and recall information declined with higher age; older people tended to perform more poorly than younger adults at tasks that required retrieving information from their short-term memory. Older people compensated for these deficits with a more focused selection of situation-defining information (Radvansky 1999).

[4] Alternatively, presentations of videos or pictures can be used.

[5] Besides information overload, there could also be a high cognitive load when the information provided is low, because respondents have to add further (not provided) information to evaluate the vignettes.

in handling the response task: *response times* and *response consistency*. Additionally, the study tests if both indicators change in the course of the evaluation, which would indicate learning and fatigue effects.

Previous research has already shown that longer response times (response latencies) are well suited to identifying problematic questions, such as questions that were difficult to understand or answer (Bassilli and Scott 1996) or to signal low quality data, such as response behavior prone to context effects and satisficing (Malhotra 2008). Similarly, previous research demonstrated that high complexity in decision tasks (e.g. operationalized by a high number of attributes in the described decisions and a high number of decision tasks presented to single respondents) and a low cognitive memory capacity of respondents (operationalized by their higher age) lead to comprehension errors and inconsistent responses (see for example Caussade et al. 2005; Finucane et al. 2005; Swait and Adamowicz 2001).

Research on the effects of the complexity of vignettes is still lacking (for the only exception: Auspurg, Hinz and Liebig 2009). To the best of our knowledge, previous applications of FSs used numbers of dimensions ranging from three (Berk and Rossi 1977) up to 21 (Shlay et al. 2005). Similarly, the applied number of vignettes presented to single respondents showed a high heterogeneity, ranging from one (e.g. Jann 2005) to 110 vignettes (Bose and Rossi 1983). None of these studies has reported a cognitive overload or symptoms for incorrect answers. However, the data quality was not explicitly considered. Since research on the working memory suggests that respondents can keep only about seven pieces of information in their short-term memories at any one time (Miller 1956), it has to be expected that higher numbers of vignette dimensions cause an overload of the memory capacity and therefore lead to extraordinarily long response times or inconsistent responses. Hence, in this study one has to expect longer response times and more inconsistent response in the experimental split with 12 vignette dimensions compared to those with only five or eight vignette dimensions.

Even though the impact of the number of vignettes on response times and response consistency has not received any attention for FSs (for the only exceptions: Auspurg, Hinz and Liebig 2009; Sauer et al. 2009), the effects on consistency have been studied in related survey methods like conjoint analysis and choice experiments. Both methods resemble FSs in that respondents are asked to evaluate several short descriptions of objects or situations, consisting of dimensions that vary in their levels (for more details: Louviere, Hensher and Swait 2000; Orme 2006). Research on these methods has shown that the response consistency, as measured by the level of unexplained variance of the lead, was higher when more attributes were varied within the choice tasks, and also markedly declined after about the tenth evaluation questioned from single respondents (see for example: Bradley and Daly 1994; Caussade et al. 2005; Swait and Adamowicz 2001). Therefore, one has to assume that more than ten vignettes per respondent lead to longer response times and lower response consistency, especially when using more than eight dimensions within the vignettes.

This should be true especially when respondents have low working memory capacities. Previous research has already shown that older and less well-educated respondents tended to need more time than younger respondents to finish surveys (Schwarz et al. 1999; Yan and Tourangeau 2008) or produced more response errors, such as responses prone to context effects (e.g. Knäuper 1999). Furthermore, older respondents were shown to produce less consistent responses than younger ones (Finucane et al. 2005). Therefore, a similar pattern is assumed here: older or less well-educated respondents need more time to evaluate vignettes and also evaluate them less consistently than younger or more educated respondents. Both assumptions should again especially apply to complex settings (with a high number of dimensions and a high number of vignettes). The same conclusions arise from the literature on bounded rationality (Simon 1955) and satisficing response behavior (Krosnick 1991).

Furthermore, the repeated evaluation task may lead to learning or fatigue effects (see for FSs: Auspurg, Hinz and Liebig 2009; for conjoint analyses and choice experiments: e.g. Bradley and Daly 1994 ; Caussade et al. 2005; Johnson, Lehmann and Horne 1990). On the one hand, learning effects most probably occur after some vignettes have been evaluated; on the other hand, fatigue might evoke a lower processing motivation and capacity and therefore increase response errors and produce a declining consistency of responses in the course of evaluating. Auspurg et al. (2009; using a student sample) found a curvilinear, inverted u-shaped connection between consistency and vignette position fitting the assumption of a learning effect that is increasingly displaced by fatigue effects. The present study therefore assumes a curvilinear connection of both response time and consistency with vignette position.

3 Methods

3.1 Study Overview

The data in this study were collected from a general population survey conducted to study method issues in FS design (n=1,634 respondents). The target population comprised inhabitants of Germany with a minimum age of 18 years. The thematic topic was the fairness of earnings (Alves and Rossi 1978). Respondents had to evaluate the gross earnings of fictitious employees. In order to study the differences in the cognitive demand of FS modules, the implemented FS varied by the number of dimensions (five, eight, 12) and the number of vignettes (10, 20, 30), yielding a 3x3 experimental setup. Hence, there was a rich variation of cognitive demand from 10 vignettes with five dimensions up to 30 vignettes with 12 dimensions. In addition, and independently of this experimental setup, the survey mode was divided into three categories: computer assisted personal interviews (CAPI), computer assisted self-interviews (CASI), and paper and pencil interviews (PAPI).[6]

[6] Within each mode of questioning the respondents were randomly assigned to one of the nine experimental splits (five, eight, or 12 dimensions x 10, 20 or 30 vignettes).

Data gathering was based on two random samples: (1) The CAPI study used a random route sample[7] with 129 sample points in Germany to select private households and a random mechanism (a Kish-selection grid) to select a target person within each household. The response rate of the random route sample was 48%[8]: 777 out of 1,608 randomly selected people participated in the study.[9] (2) The sample for CASI and PAPI was drawn by using a random digit dialing method modified for Germany (Häder and Gabler 1998). For this sample, private households were contacted over their landline telephone number. A random mechanism (a Kish-selection grid) provided the target participants within the households. These target participants had to answer a few screening questions on their willingness to participate and to indicate their e-mail or postal addresses. They could choose between a CASI and a PAPI version of the questionnaire. The CASI version was sent by e-mail link and the PAPI version by regular mail (with an envelope for returning the questionnaire). In this sample, the return rate for the 1,423 people successfully recruited (who provided an e-mail address or a mailing address) was 60%; 461 people participated in the online survey and 396 people returned the printed questionnaire.[10]

For the CASI and CAPI version, para-data on response times were collected for each vignette (time stamps). In order to standardize the CAPI and CASI mode, the interviewers in the CAPI mode passed the laptop to the respondents when the FS module started. Respondents were asked to read and evaluate the vignettes on their own (with the interviewers providing assistance if needed). Table 1 gives an overview of the observations for each of the nine conditions to which respondents were randomly assigned.[11] The FS module implemented focused on the evaluations of fairness in earnings of fictitious employees who were described using five, eight or 12 dimensions.[12] All fictitious employees were characterized as working full-time (40 hours per week). The respondents had to rate 10, 20 or 30 vignettes each by using an 11-point-scale ranging from - 5 ("far too low"), to 0 ("fair") and + 5 ("far too high").[13]

3.2 Vignette Dimensions and Levels

The chosen dimensions presented on the vignettes were based on previous evidence from FSs on the fairness of earnings (Alves and Rossi 1978; Jasso 1978; Jasso and Rossi 1977; Jasso and Webster 1997, 1999). According to this research, the vignette-variables age, sex, number of children, occupational prestige (as measured by SIOPS)[14] and educational degree were expected to have an impact on the fairness evaluations. Further dimensions known to be relevant from previous fairness studies and labor market research were also added; these dimensions comprised job performance and marital status (Struck, Krause and Pfeifer 2008). Some contextual characteristics regarding the work organization of the fictitious employees were added. Table 2 shows the used dimensions and levels (for exemplary of vignettes with five, eight and 12 dimensions, see Table A1 in the Appendix).

3.3 Vignette Sample

The vignettes were selected via a quota design (*D-efficient design*) to preserve the preferable features of the vignette universe (where all dimensions are mutually uncorrelated and all levels occur with the same frequency).

D-efficient designs are built using a computer algorithm that specifies a sample characterized by a minimal intercorrelation between all or the most important dimensions (and interaction terms) and at the same time a maximal variance and balance of the frequency of the vignette levels. These designs ensure that the influence of vignette dimensions and interaction terms are mutually uncorrelated (Atzmüller and Steiner 2010; Dülmer 2007). In addition, the design features lead to minimal standard errors in data analyses and therefore a higher statistical "power" and efficiency (in a statistical sense) to reveal the influence of single dimensions than other designs (like random samples).

For the study at hand, a D-efficient sample of 240 vignettes was build up (illogical cases, e.g., medical doctors without a university degree, were excluded).[15] The employed sample of vignettes with 12 dimensions reached a D-

[7] In a random route design, each interviewer receives a randomly assigned starting address per sampling point (usually electoral districts) and moves along from that address by a script with specific instructions. These instructions follow an algorithm indicating the streets, houses, and apartments to choose. In Germany, this method is regularly used to draw random samples of private households (ADM 1999). See Schnell, Hill and Esser (2008) for a textbook description.

[8] The response rate is calculated according to American Association of Public Opinion Research standards (AAPOR 2011).

[9] The survey was conducted by the survey institute USUMA (Berlin, Germany).

[10] The survey institute reported a total number of 3,308 persons contacted; 43% of them provided an e-mail or postal address.

[11] The distribution of respondents to the experimental splits was independent of the mode of questioning (*likelihood-ratio* χ^2 = 7.470, df = 16, p = .963). In addition, there was no correlation of respondents' characteristics like age or educational group with experimental splits (ANOVA with age and experimental group: F= 0.821, p= .584; likelihood-ratio test on educational degree and experimental group: *likelihood-ratio* χ^2 = 12.841, df = 16, p = .684). The number of dimensions was not correlated with the number of vignettes (*likelihood-ratio* χ^2 = 0.203, df = 4, p =.995).

[12] All vignettes for each respondent had the same number of dimensions (*between-subject* design).

[13] Only 17 respondents (1.05%) dropped out during the vignette module. All of them were respondents in the CASI mode. 375 vignettes (1.39%) were not evaluated by the respondents. The settings with most missing values were the split with 12 dimensions (n= 182; 48.5% of all missing values) and the split with 30 vignettes (n= 161; 42.9% of all missing values).

[14] Standard International Occupational Prestige Scale (Ganzeboom and Treiman 1996).

[15] The exclusion of illogical cases is recommended because one preferable feature of vignettes is that they describe realistic scenarios. Illogical cases can lead to undesirable method effects like fading out of dimensions by the respondents (Auspurg, Hinz and Liebig 2009).

Table 1: Study setup and number of cases (respondents)

	CAPI			CASI			PAPI		
Vignettes	5 dim.	8 dim.	12 dim.	5 dim.	8 dim.	12 dim.	5 dim.	8 dim.	12 dim.
10	132	136	147	85	85	78	72	70	76
20	65	73	75	46	45	34	34	28	39
30	51	52	46	29	28	31	27	23	27

Table 2: Vignette dimensions and their levels

#	Dimensions	Levels
1	Age	30/40/50/60 years
2	Sex	Male/female
3	Vocational degree	Without degree/vocational degree/university degree
4	Occupation	Unskilled worker/door(wo)man/engine driver/clerk/ hairdresser/social worker/software engineer/ electrical engineer/manager/medical doctor
5	Gross earnings per month	Ten values ranging from 500 to 15,000 Euros
6	Experience	Short on/much
7	Job tenure	Entered recently/entered a long time ago
8	Children[a]	No child/1 child/2 children/3 children/4 children
9	Health status[b]	No health problems/health problems for a longer time
10	Performance	Under/above average
11	Economic situation of the firm	High profits/threatened by bankruptcy/solid
12	Firm size	Small/medium/large enterprise

[a] The category *no child* was oversampled to achieve a more realistic distribution of family size.

[b] The category *no health problems* was oversampled for a similar reason.

efficiency of 90.7; for more details on the implementation in the project at hand: Auspurg and Wehrli 2009; Sauer et al. 2009).[16] For the splits with five and eight dimensions, all irrelevant dimensions were deleted (these were the dimensions 6 to 12 respectively 9 to 12 in Table 2). Although it may have been possible to achieve more efficient samples for each split individually (samples with lower correlations and a higher variance for all dimensions), the focus was to hold the statistical efficiency constant over the different splits with more or fewer dimensions or more and fewer vignettes in order to focus on the pure impact of these method aspects (Auspurg, Hinz and Liebig 2009). This procedure guaranteed that correlations and variances of the vignette dimensions were the same for all experimental splits. Selections of 10, 20 or 30 vignettes and the sequence of vignettes were randomly assigned to each respondent.

3.4 Analysis

The first part of data analysis focuses on response time. Time stamps allow a calculation of the time span respondents needed to read and evaluate the vignette presented. This time span is called the raw reaction time (RT).[17] Respondents usually differ to a great extent in their baseline speed (BS) because of different experience levels with computers and web surveys or because they are generally faster or slower with reading questions and responding. Therefore, the analyses are based on a transformed RT, the latency time (LT), which is uncorrelated with BS. LT is not a general measure of speed but reveals whether or not the respondents needed a particularly long time to answer the vignettes, which might signal a particularly high cognitive load for the task. Since the LTs are adjusted for the BS, longer response times indicate a cognitive load that surpasses the known differences between age and educational groups in processing times (which are already captured in the BS). To measure the LTs in the study at hand, the following procedure was used. First, the BS for each respondent was estimated using a different part of the survey (a four-question knowledge test which was self-administered like the vignette module in all survey modes and which was placed at the end of the questionnaire). Secondly, the RT was adjusted for the vignettes on BS using a residual-score-index approach (Mayerl and Urban 2008; Urban and Mayerl 2007). The residuals from a regression of the RT on the BS were used as LTs;[18] as Mayerl and Urban (2008) have shown, this approach leads to more accurate estimates of LTs than alternative methods.

[16] The D-efficiency is a measure of the goodness of designs that captures both the orthogonality and variance of dimensions. It is scaled from 0 to 100; values from 90 up can in general be considered to be adequate (for more details: Dülmer 2007).

[17] 700 time measures were identified as outliers (out of 17,215 observations) with RTs of 127 seconds or above. The procedure of outlier detection for the high values was as follows. First, those values were discarded above the 99th percentile; secondly, those measures were dropped that were two standard deviations above the mean (for a recommendation of this procedure: Mayerl and Urban 2008).

[18] That is, the LTs are the residuals δ_{ij} from the following regression: $RT_{ij} = a_0 + b_1 BS_j + \delta_{ij}$; with the subscript i indicating the single vignette (up to 30 for each respondent) and j indicating the single respondent.

The second part of the analysis dealt with a straightforward measure of consistency. Following similar approaches in the research into the response consistency in choice experiments (Caussade et al. 2005; Swait and Adamowicz 2001), the analysis was based on the unexplained variance in the vignette judgments. More concretely, the inconsistency of responses was measured by the squared level-one residuals from random intercept regressions with the vignette rating serving as dependent variable and all vignette dimensions as predictors.[19] Afterwards, the squared residuals out of this regression were used as a measurement of response consistency. The higher the consistency of responses, the smaller the amount of variance unexplained through the vignette dimensions and the smaller the squared residuals.

For both indicators of cognitive load, the response times and response consistency, it was analyzed whether they were (given the varying complexities of the vignettes in the experimental splits) systematically related to respondents' characteristics (educational background and age). For the response consistency, it was also of interest if it had declined with the complexity of vignettes (the number of dimensions and the number of vignettes). In addition, it was investigated if both measures for a high cognitive load changed in the course of the evaluations, which would indicate learning or fatigue effects.

For the statistical analyses, respondents were categorized into three age groups (18 to 41 years, 42 to 59 years, and 60 years and older; each group representing around one third of respondents) and three educational groups (general educational level: lower, middle, and higher secondary school certificate).[20] Because the level of difficulty of the fairness evaluations for respondents could be related to their labor market experience and their knowledge of wage structures, and because both experience and knowledge with the question topic are known to reduce cognitive load for respondents (e.g. Tourangeau, Rips and Rasinski 2000), also a control variable for current labor market status was included (dichotomized: 1 for "employed" , 0 for "not employed").[21]

To analyze both dependent variables (response times and response consistency), a multilevel model (random intercept model) was used that accounted for multiple ratings per respondent (Hox, Kreft and Hermkens 1991). In random intercept models, there are two error components used to build up the hierarchical data structure that results from several observations belonging to each single respondent. The error component on level two, often referred to as the random intercept, measures the between-respondent variation, while an additional error component on level one captures the variation within the vignettes judgments of single respondents.[22]

If the complexity of the evaluation task is not relevant for LTs or consistency, *none* of the design variables (the number of vignettes and the number of dimensions) or of the respondents' characteristics, indicating their cognitive abilities, should be systematically related to the LTs or squared residuals. In this case, the model fit (the χ^2-value) of the multilevel regressions would be insignificant, meaning that the H_0 – the assumption that the R^2-value of the regression is zero in the population – cannot be rejected. In other words, the model would have no explanatory power at all. Only in the other cases, where at least one of the design variables or respondents' characteristics are systematically related to the measures of cognitive load (LTs and response consistency), is their impact discussed in more detail.

For both dependent variables, separate regressions for different numbers of dimensions served to test if the impact of the cognitive ability of respondents is dependent on the complexity of the vignettes. Because a low internal consistency of responses would in particular signal a problematic response behavior, for the analyses of response consistency, the estimates are additionally presented separately for the conditions with 10, 20, and 30 vignettes. These analyses make it possible to test if – as expected – the measures of complexity moderate each other and if, in particular, the most complex condition (30 vignettes with 12 dimensions) causes respondents with low cognitive abilities (high age, low education) to produce a low consistency of responses.[23]

4 Results

4.1 Response Time

The overall time for the FS module varies by complexity. For 10 vignettes and five dimensions the median overall RT of the module was 2.85 minutes. The most complex setup, with 30 vignettes and 12 dimensions, had a median overall RT of 9.55 minutes (for results on all settings see Table A2 in the Appendix). These descriptive statistics show that, even

[19] The following regression model was used:

$J_{ij} = a_0 + b_1 v\_sex_{ij} + b_2 v\_age_{ij} + b_3 v\_voc_{ij} + b_4 v\_uni_{ij} + b_5 v\_siops_{ij} + b_6 v\_earn_{ij} + u_j + \varepsilon_{ij}$; with J_{ij} denoting the vignette rating of respondent j for vignette i.

[20] The three educational groups were constructed according to the three tracks distinguished within the German educational system. The usage of the categorical age groups allows plotting effects of the treatments separately for respondents with different age. The results differed only marginally when using metric representations of age.

[21] "Not employed" respondents included those who were unemployed as well as those who were not looking for employment or were retired at the time of the survey.

[22] The random intercept models were appropriate because there was no correlation between the respondent specific error component u_j and the independent variables (the correlation between u_j and the covariates was r = -.0004). The assumption of independence was tested for all multi-level models presented hereafter. In the CAPI mode, the data structure had a third hierarchical level because the interviews were nested for interviewers. Interviewer effects might have occurred and have influenced the preferences of respondents in terms of their fairness evaluations. However, separate analyses (not shown here) demonstrated that the results were robust to interviewer effects. Therefore, the sparse specification of only two levels (respondents and vignettes) is shown here.

[23] Within all models on the response consistency, the dependent variable consisted of squared residuals out of the regression model from the equation in footnote 19. That is, the same dimensions were always used. This procedure allowed the standardization of the degrees of freedom and substantive dimensions across all regression estimates.

for the most complex FS module, most respondents were able to answer the vignette module in less than ten minutes, which might be a tolerable amount of time for most practical applications.

To gain an impression of how RT depends on the vignette positions (1 up to 30) and complexity of vignettes (number of dimensions), the development of RTs over the evaluation course for those respondents who rated 30 vignettes is shown in Figure 1 (row 1). In the graphs for the five-, eight-, and 12 dimensions conditions, the patterns are similar. The respondents needed more time to rate the first vignettes but then responded faster on a stable level. Figure 1, row 2, shows the RT for the most complex vignette module (30 vignettes with 12 dimensions) by age group. For all three age groups, there was a sharp decline of RT within the first part of the vignette module. Older respondents required more time than younger respondents, especially when evaluating the first few vignettes with a high number of dimensions. For further vignettes, RT was nearly constant in all age groups.

However, these age effects might be not special for FSs since older respondents in general need longer to read texts or to answer survey questions. It is more interesting therefore if older and less educated respondents need a particularly long time to evaluate the vignettes. This can be answered by studying the LTs per vignette, which are, as already explained, standardized for the individual BS. Table 3 provides the coefficients of the multilevel regression for LTs, separately for the split with five, eight and 12 dimensions.

All χ^2-values indicating the model fit were significantly different from zero, meaning that all three regression models had explanatory power. Yet there were only two statistically significant design effects. In both the splits with five and eight dimensions, the CAPI mode was characterized by a longer LT (on the average, 2.4 respectively 1.7 seconds more per vignette). This could reflect a self-selection into the CASI mode which was not fully controlled by considering the BS. The presence of an interviewer might also have changed the information processing. Contrary to stated expectations, there were no significant effects for respondents' characteristics. Even in the split with the most complex vignettes (12 dimensions), respondents' age or education were not correlated with the LTs. Altogether, the analyses on response times suggest that there is a learning effect, but do not reveal a higher cognitive load or effort for respondents who are typically known for less cognitive memory capacities.

4.2 Consistency of Evaluations

The analysis in Table 4 addresses possible effects of the complexity of vignettes by estimating separate regressions for the five, eight, and 12 dimensions conditions. Is there any evidence for specific factors causing a higher inconsistency of responses, especially when vignettes are described by more dimensions? The overall fit of the first two models (five and eight dimensions) was insignificant, and response consistency was therefore not influenced by the design and respondents' characteristics in both these models.[24] The third model (12 dimensions) showed a significant χ^2-value, which was mainly driven by a CASI effect. Those respondents who evaluated complex vignettes in this mode tended to produce higher squared residuals. A more detailed analysis of this CASI effect showed that it was especially strong for those respondents who were not employed at the time of the survey. For those individuals, the substantive topic of the survey could be slightly more demanding and maybe led to an overload, especially when there was no interviewer-assistance. Furthermore, there was a weak but insignificant tendency that respondents with middle or high education produced a higher consistency (smaller residuals) than other respondents; this was also true for the middle age group.

Finally, Table 5 shows the results of three multivariate models predicting the residuals for the 10, 20 and 30 vignettes condition. The overall fit of models 1 (10 vignettes) and 2 (20 vignettes) indicated that the regressors measuring the design features or respondents' cognitive abilities in sum had very low (10 vignettes) or no explanatory power at all (20 vignettes; see the insignificant χ^2-Value). In other words, especially for the split with 20 vignettes (Model 2), the consistency was not influenced by features of the study like mode of administration or number of dimensions and moreover was not affected by respondents' age or educational level.

The results were different for the 30 vignettes condition shown in Model 3. In this condition, the vignettes with 12 dimensions showed higher residuals compared to the reference group of five dimensions. Moreover, the effect of the vignette position on the residuals was in accordance with a curvilinear, s-shaped pattern (see the positive coefficients for the vignette position and vignette position ^3, and the negative coefficient for the vignette position ^2). The residuals became smaller during the first part of the FS module, reaching a minimum at vignette position 12. The residuals then increased until vignette 23. This pattern corresponds to both the expected learning and fatigue effects.

In addition, it seems that the respondents concentrated more during the final part of vignettes: in this part, the residuals shrunk again, as was already found in a prior vignette study (Auspurg et al. 2009b). In terms of the individual variables, Model 3 is characterized by a significant impact of respondents' education. Respondents of the middle and higher educational group produced a smaller amount of residuals compared to respondents with a lower level of education.

To sum up, the analyses on the response consistency mainly indicate that only in cases of very complex FSs (more than eight dimensions and more than 20 vignettes per respondent) respondents, in particular those with low educational level, produce less consistent answers, which signales cognitive overload. Additionally, only in cases with such a complex survey design there was evidence for learning and fatigue effects and for contending that the experience of respondents with the evaluation task matters for the response consistency. The age of respondents, surprisingly, did not matter for the response consistency at all. All in all, the ex-

[24] Several model assumptions were tested to exclude the possibility that the low χ^2-Values originated from misspecifications of the regression equation.

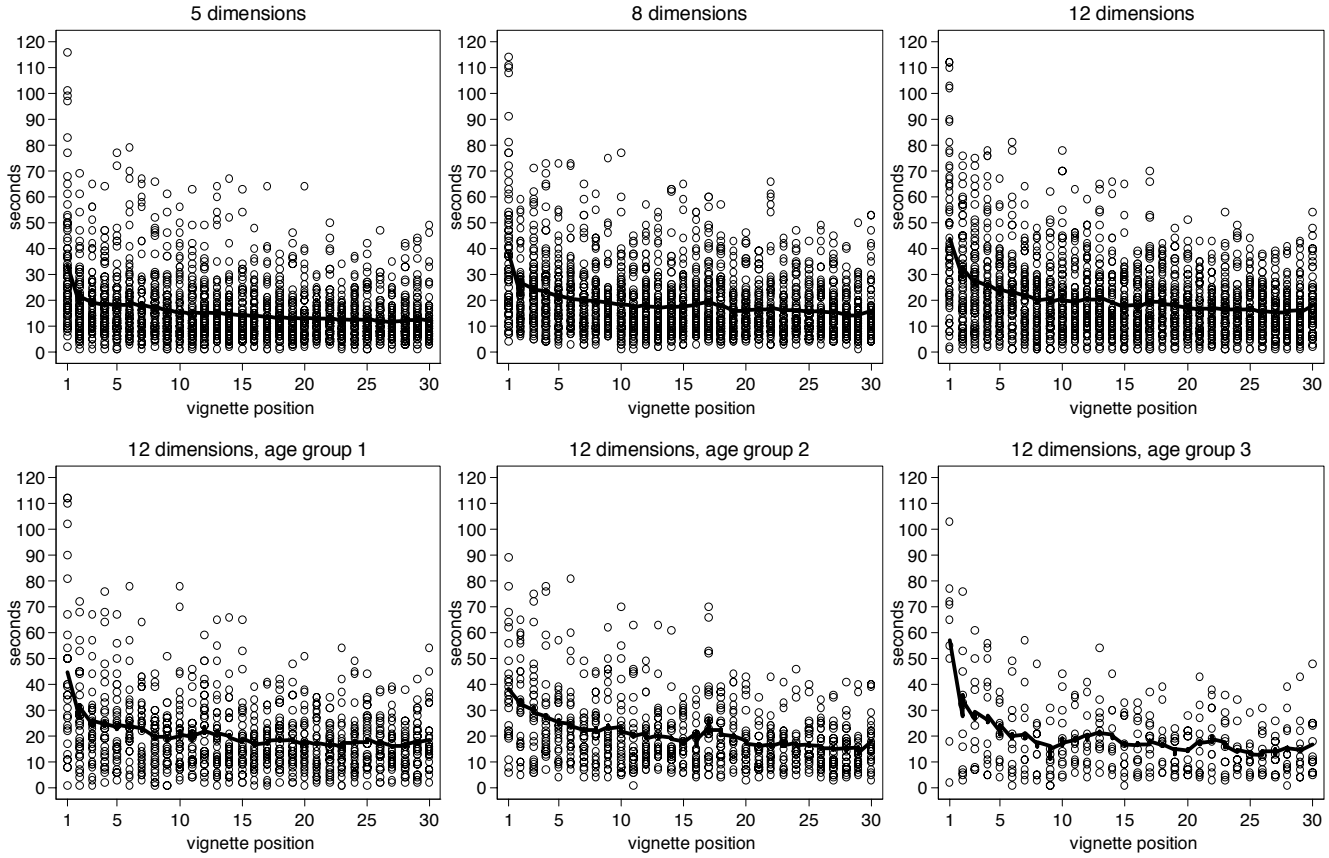

Figure 1. Raw reaction time (RT) by vignette position, dimensions and age groups [in seconds]

Table 3: Latency time, age and educational background

	5 dimensions		8 dimensions		12 dimensions	
	b	se	b	se	b	se
CAPI	2.359**	.826	1.739*	.877	.732	.986
Lower sec. school	ref.		ref.		ref.	
Middle sec school	-.259	.889	.270	.963	.956	1.018
Higher sec. school	.947	1.042	.479	1.131	.090	1.209
18-41 years	ref.		ref.		ref.	
42-59 years	1.197	.864	1.759	.935	.826	1.006
60 years and older	1.612	1.032	1.713	1.086	-1.561	1.256
Employed [1 = yes]	-.534	.824	.635	.885	-1.467	1.008
Constant	9.325***	1.218	13.172***	1.319	19.496***	1.484
Sigma_u	6.200		6.568		7.292	
Sigma_e	10.367		11.026		11.501	
Vignettes	5408		5698		5409	
Respondents	341		353		356	
Wald Chi^2(df=35)	850		808		963	
p-value	<.001		<.001		<.001	
R^2	.123		.109		.139	

Generalized least squares (GLS) regression models; dependent variable (DV): latency time, controlled for vignette position.* p <.05, ** p <.01, *** p <.001

Table 4: Regression on squared residuals by vignette dimensions

	5 dimensions		8 dimensions		12 dimensions	
	b	se	b	se	b	se
10 vignettes	ref.		ref.		ref.	
20 vignettes	-.230	.173	-.111	.151	.030	.170
30 vignettes	-.190	.180	.168	.160	.274	.180
CAPI	ref.		ref.		ref.	
CASI	-.179	.182	.137	.160	.598**	.186
PAPI	-.052	.192	.176	.176	.243	.190
Lower sec. school	ref.		ref.		ref.	
middle sec school	.025	.175	-.013	.160	-.292	.170
higher sec. school	-.137	.206	.054	.182	-.282	.197
18-41 years	ref.		ref.		ref.	
42-59 years	.298	.182	.085	.166	-.344	.181
60 years and older	.050	.205	-.191	.186	-.268	.218
Employed [1 = yes]	-.076	.171	-.089	.155	-.170	.178
Constant	2.620***	.210	2.388***	.193	2.772***	.209
Sigma_u	1.220		1.011		1.195	
Sigma_e	4.358		4.121		4.385	
Vignettes	8669		8642		8674	
Respondents	527		526		537	
Wald Chi^2 (df=9)	7.535		6.389		22.651	
p-value	.582		.700		.007	
R^2	.017		.011		.034	

GLS regression models; DV: squared residuals
$^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001

Table 5: Regression on squared residuals by number of vignette

	10 vignettes		20 vignettes		30 vignettes	
	b	se	b	se	b	se
5 dimensions	ref.		ref.		ref.	
8 dimensions	-.209	.153	-.115	.160	.170	.198
12 dimensions	-.084	.151	.137	.160	.391*	.198
Vignette position	.267	.209	-.059	.089	-.162**	.059
Vignette $position^2$ [*10]	-.737	.431	.022	.098	.105*	.044
Vignette $position^3$ [*10]	.048	.026	-.000	.003	-.002*	.001
CAPI	ref.		ref.		ref.	
CASI	.165	.157	.045	.164	.390	.203
PAPI	.184	.161	-.194	.178	.414	.218
Lower sec. school	ref.		ref.		ref.	
Middle sec school	.301*	.150	-.258	.155	-.552**	.197
Higher sec. school	.110	.170	-.088	.187	-.494*	.227
18-41 years	ref.		ref.		ref.	
42-59 years	.166	.155	-.179	.163	-.110	.212
60 years and older	-.094	.179	-.247	.193	-.183	.233
Employed [1 = yes]	-.161	.148	-.051	.156	-.112	.197
Constant	2.340***	.336	3.076***	.294	3.472***	.317
Sigma_u	1.228		.952		1.153	
Sigma_e	4.183		4.154		4.488	
Vignettes	8485		8382		9118	
Respondents	857		425		308	
Wald Chi^2(df=12)	21.965		18.416		44.654	
p-value	.038		.104		<.001	
R^2	.015		.021		.054	

GLS regression models; DV: squared residuals
$^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001;

plained variance of the models and therefore the indications for cognitive overload were low.

5 Discussion

The FS approach is considered an appropriate instrument for measuring normative judgments, social attitudes or hypothetical decisions rules. So far, however, there is limited knowledge about method effects when FSs are applied to heterogeneous population samples. This study has focused on how respondent characteristics associated with cognitive abilities (respondents' age and education) interact with the complexity of the evaluation task in a FS module. The complexity was varied by the number of vignettes and their number of dimensions. As dependent variables, both the response times (measured by RTs and LTs) and the response consistency (measured by the unexplained variance of vignette judgments) were investigated.

All in all, the main findings support the applicability of the FS approach in general population surveys. First, even given a high number of vignettes and dimensions, the RTs were not particularly long. The FS module took on average between 2.85 and 9.55 minutes depending on the complexity of the design. The development of RTs over vignette positions was characterized by a decline after the first vignettes; obviously a few vignettes were needed to learn the rating task at hand. Afterwards, the RT per vignette stayed almost stable. Second, even if older respondents showed longer RTs, the analyses of the LTs (RTs that were adjusted for the individual BS) found that there was no impact of either age or educational background. Regardless of the measured variation of cognitive ability, the vignette task did not cause lengthened response times. Third, the analyses of the response consistency showed almost no effects of the vignette position, mode or respondents' characteristics for the less complex set-ups with no more than eight vignette dimensions and no more than 20 vignettes per respondents. Only in cases of very complex conditions (12 dimensions or 30 vignettes) did design variables, like the survey mode, or respondent characteristics, like education, have any explanatory power for response consistency.

For the 30-vignette condition, there was a curvilinear effect of vignette order; this is in line with possible learning and fatigue effects accompanied by a potentially higher concentration for the last vignettes evaluated. Additionally, for the 30-vignette condition, the most complex design with 12 dimensions led to a higher inconsistency of responses, indicating information overload. Furthermore, there were effects of the educational level, showing that respondents with a lower secondary degree produce a larger amount of inconsistency within their ratings than respondents with higher levels of education.

For the 12-dimension condition, one single method effect reached statistical significance: there was a lower consistency of response in the CASI mode. The complex vignettes with long text were maybe exhausting or more difficult to read on the computer screen.[25] Surprisingly, the age of respondents was not correlated to response consistency at all.

Summing up, these results suggest that the FS task does not demand a higher cognitive effort, as measured in LTs, than other questions, and that it triggers differences in response consistency only across educational groups or survey modes in cases of very complex settings with more than eight dimensions and more than 20 vignettes. This is good news, since most (but not all) applications so far are based on less complex settings (for a review: Wallander 2009).

Further research should investigate whether these results are replicable and generalizable to FSs dealing with other substantive topics or respondent samples. Another recommendation is to study whether respondents fade out some dimensions of the vignettes to reach consistent evaluations. There is some evidence from research on related methods like choice experiments and conjoint analyses that respondents use a lot of information when evaluating the first vignettes and concentrate only on a number of more salient dimensions later on (Bradley and Daly 1994; Caussade et al. 2005). If these results were true for FSs as well, further research would be needed to find out if those response heuristics lead to valid responses and mirror judgment or decision behavior in real situations (Gigerenzer and Todd 1999) or if those heuristics on the contrary represent a problematic response behavior caused by the artificial, repeated evaluation task in FSs. In addition, there is more research needed to understand the mode effect and to test how the cognitive load, and so response consistency, is related to respondents' knowledge and experience with the evaluation task.

What can nevertheless be learned from the study at hand is that the implementation of FSs in general population samples should take cognitive restrictions into careful consideration; one recommendation would be to limit the number of vignettes to no more than 20 vignettes with fewer than 12 dimensions in order to avoid fatigue effects and inconsistent responses. The general message of our paper is, however, that FSs are applicable in general population surveys and that respondents with different age and educational backgrounds are able to perform the demanding evaluation tasks of FSs and produce a high level of response consistency during ordinary response times.

Acknowledgements

We thank two anonymous reviewers and one editor of this journal for many useful comments and suggestions. All shortcomings are the responsibility of the authors.

References

ADM. (1999). *Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis*. Germany, Opladen: Leske + Budrich.

Alves, W. M., & Rossi, P. H. (1978). Who should get what? Fairness judgments of the distribution of earnings. *American Journal of Sociology*, *84*(3), 541-564.

[25] Since this CASI effect was especially strong for those respondents who did not participate in the labor market, the assistance of an interviewer might be especially helpful for respondents that are less familiar with the substantive topic addressed by the FS module.

American Association for Public Opinion Research (AAPOR). (2011). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (7th ed.). Deerfield, IL: AAPOR.

Atzmüller, C., & Steiner, P. M. (2010). Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128-138.

Auspurg, K., Hinz, T., & Liebig, S. (2009). Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. *Methoden Daten Analysen*, *3*(1), 59-96.

Auspurg, K., Hinz, T., Sauer, C., & Liebig, S. (2009). *How Numbers Matter. Experimental Results on the Effects of Complex Settings in Factorial Survey Designs.* Paper presented at the 3rd PPSM Meeting, Bremen [May 7th].

Auspurg, K., & Wehrli, S. (2009). *Codebuch und Dokumentation. Einkommensgerechtigkeit in Deutschland.* Technical Report # 1 of the DFG-Project 'The Factorial Survey as a Method for Measuring Attitudes in Population Surveys'. ETH Zurich / University of Konstanz.

Bassilli, J. N., & Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly*, *60*, 390-399.

Berk, R. A., & Rossi, P. H. (1977). *Prison reform and state elites*. Cambridge, Mass: Ballinger.

Bose, C. E., & Rossi, P. H. (1983). Gender and jobs: prestige standings of occupations as affected by gender. *American Sociological Review*, *48*, 316-330.

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, *21*(2), 167-184.

Brody, N. (1997). Intelligence, schooling, and society. *American Psychologist*, *52*(10), 1046-1050.

Caussade, S., Ortúzar, J., Rizzi, L. I., & Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation research part B: Methodological*, *39*(7), 621-640.

Copeland, D. E., & Radvanksy, G. A. (2007). Aging and integrating spatial mental models. *Psychology & Aging*, *22*(3), 569-579.

Dülmer, H. (2007). Experimental Plans in Factorial Surveys: Random or Quota Design? *Sociological Methods & Research*, *35*, 382-409.

Falch, T., & Sandgren, S. (2006). *The effect of education on cognitive ability.* Working Paper Series 7306, Department of Economics, Norwegian University of Science and Technology.

Finucane, M. L., Slovic, P., & Schmidt, E. S. (2005). Task Complexity and Older Adults' Decision-Making Competence. *Psychology and Aging*, *20*, 71-84.

Ganzeboom, H., & Treiman, D. (1996). International Comparable Measures of Occupational Status for the 1988 International Standard Classification of Occupations. *Social Science Research*, *25*, 201-239.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Häder, S., & Gabler, S. (1998). Ein neues Stichprobendesign für telefonische Umfragen in Deutschland. In S. Gabler, S. Häder, & J. H. P. Hoffmeyer-Zlotnik (Eds.), *Telefonstichproben in Deutschland* (p. 69-88). Opladen: Westdeutscher Verlag.

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Hox, J. J., Kreft, I. G., & Hermkens, P. L. J. (1991). The Analysis of Factorial Surveys. *Sociological Methods & Research*, *19*, 493-510.

Jann, B. (2005). Lohngerechtigkeit und Geschlechterdiskriminierung. Evidenz aus einem Vignetten-Experiment. In B. Jann (Ed.), *Erwerbsarbeit, Einkommen und Geschlecht. Studien zum Schweizer Arbeitsmarkt* (p. 107-126). Wiesbaden: VS Verlag für Sozialwissenschaften.

Jasso, G. (1978). On the justice of earnings: A new specification of the justice evaluation function. *American Journal of Sociology*, *83*(6), 1398-1419.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research*, *34*(3), 334-423.

Jasso, G., & Meyersson Milgrom, E. M. (2008). Distributive Justice and CEO Compensation. *Acta Sociologica*, *51*, 123-143.

Jasso, G., & Rossi, P. H. (1977). Distributive justice and earned income. *American Sociological Review*, *42*(4), 639-651.

Jasso, G., & Webster, M. J. (1997). Double standards in just earnings for male and female workers. *Social Psychology Quarterly*, *60*(1), 66-78.

Jasso, G., & Webster, M. J. (1999). Assessing the gender gap in just earnings and its underlying mechanisms. *Social Psychology Quarterly*, *62*(4), 367-380.

Johnson, M. D., Lehmann, D. R., & Horne, D. R. (1990). The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing*, *7*(1), 35-43.

Knäuper, B. (1999). The Impact of Age and Education on Response Order Effects in Attitude Measurement. *The Public Opinion Quarterly*, *63*, 347-370.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *The Public Opinion Quarterly*, *51*(2), 201-219.

Louviere, J., Hensher, D. A., & Swait, J. D. (2000). *Stated Choice Methods. Analysis and Application*. Cambridge: Cambridge University Press.

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly*, *72*, 914-934.

Mayerl, J., & Urban, D. (2008). *Antwortreaktionszeiten in Survey-Analysen: Messung, Auswertung und Anwendungen*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits in our capacity for processing information. *Psychological Review*, *63*, 81-97.

Narayan, S., & Krosnick, J. A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, *60*(1), 58-88.

Orme, B. K. (2006). *Getting started with conjoint analysis. Strategies for product design and pricing research*. Madison/Wisconsin: Research Publishers LLC.

Radvansky, G. A. (1999). Aging, memory, and comprehension. *Current Directions in Psychological Science*, *8*(2), 49-53.

Radvansky, G. A., & Copeland, D. E. (2004). Working memory span and situation model processing. *The American Journal of Psychology*, *117*(2), 191-213.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, & P. H. Rossi (Eds.), *Qualitative and quantitative social research: Papers in honour of Paul F. Lazarsfeld* (p. 176-186). New York: Free Press.

Rossi, P. H., & Anderson, B. (1982). The Factorial Survey Approach: An Introduction. In P. Rossi & S. L. Nock (Eds.), *Mea-*

suring social Judgements: the Factorial Survey Approach (p. 15-67). Beverly Hills: Sage.

Sauer, C., Liebig, S., Auspurg, K., Hinz, T., Donaubauer, A., & Schupp, J. (2009). *A factorial survey on the justice of earnings within the SOEP-Pretest 2008.* IZA Discussion Paper, 4663.

Schnell, R., Hill, P., & Esser, E. (2008). *Methoden der empirischen Sozialforschung* (8th ed.). München: Oldenbourg.

Schwarz, N., Park, D., Knäuper, B., & Sudman, S. (1999). *Cognition, aging, and self-reports*. Washington, DC: Psychology Press.

Shlay, A. B., Tran, H., Weinraub, M., & Harmonkatrin, M. (2005). Teasing apart the child care conundrum: A factorial survey analysis of perceptions of child care quality, fair market price and willingness to pay by low-income, African American parents. *Early Childhood Research Quarterly*, *20*, 393-416.

Simon, H. A. (1955). A Behavioural Model of Rational Choice. *Quarterly Journal of Economis*, *69*, 99-118.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in Experimental Design in Attitude Surveys. *Annual Review of Sociology*, *22*, 377-399.

Struck, O., Krause, A., & Pfeifer, C. (2008). Entlassungen: Gerechtigkeitsempfinden und Folgewirkungen. Theoretische Konzepte und empirische Ergebnisse. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *60*(1), 102-122.

Swait, J., & Adamowicz, W. (2001). Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice. *Organizational Behavior and Human Decision Processes*, *86*, 141-167.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.

Urban, D., & Mayerl, J. (2007). Antwortlatenzzeiten in der surveybasierten Verhaltensforschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *59*(4), 692-713.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505-520.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51-68.

Appendix

A1: Vignette examples

Vignette with five dimensions

A 60-year-old woman with vocational training works as a social worker.
Her monthly gross earnings total 2,500 Euros (before tax and extra charges).

Are the monthly gross earnings of this person fair or are they, from your point of view, unfairly high or low?

Unfairly low					Fair				Unfairly high	
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
❑	❑	❑	❑	❑	❑	❑	❑	❑	❑	❑

Vignette with eight dimensions

A 50-year-old woman with no vocational training has two children. She works as a clerk and has gained a lot of experience. She has worked for the company for a long time.

Her monthly gross earnings total 1,200 Euros (before tax and extra charges).

Are the monthly gross earnings of this person fair or are they, from your point of view, unfairly high or low?

Unfairly low					Fair				Unfairly high	
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
❑	❑	❑	❑	❑	❑	❑	❑	❑	❑	❑

Vignette with 12 dimensions

A 40-year-old woman with no vocational training has one child and works as an engine-driver. She has gained a lot of experience. She has had health problems for a long time.
She joined the (economically stable) large enterprise a long time ago and her performance is above average.

Her monthly gross earnings total 1,500 Euros (before tax and extra charges).

Are the monthly gross earnings of this person fair or are they from your point of view unfairly high or low?

Unfairly low					Fair				Unfairly high	
-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
❑	❑	❑	❑	❑	❑	❑	❑	❑	❑	❑

Table A2: Median raw reaction time (RT) by number of vignettes and dimensions [in minutes]

	Number of dimensions		
Number of vignettes	5	8	12
10	2.85	3.52	3.85
20	4.97	5.72	6.33
30	6.87	8.37	9.55